Recommendations for Evaluating Effectiveness

Executive Committee Working Group

First version: February 21, 2000
Revised and approved by the Executive Committee, September 25, 2002 [Nov. 1, 2002]

1. PREFACE

The Health Care Financing Administration (HCFA) convened the Medicare Coverage Advisory Committee (MCAC) to provide advice on scientific and clinical questions regarding coverage. In its initial charter, MCAC had an Executive Committee and six panels. Each panel addressed a different category of medical intervention. The purpose of this Executive Committee document is to provide guidance to the six panels. The goals of this document are to promote *consistency* (within and between panels) in the reasoning that leads us to a conclusion about the evidence and *accountability* (to each other and to the public) to explain our reasoning.

Each panel of the Medicare Coverage Advisory Committee will use criteria and procedures to evaluate the adequacy of the evidence and the magnitude of clinical benefit in determining the effectiveness of new medical products and services (laboratory test, diagnostic procedure, preventive intervention, treatment). This document has two purposes:

> First, it provides general guidance to the panels in the form of suggestions about how to evaluate evidence. This document makes the distinction between *adequacy* of evidence and the *magnitude* of the benefit. The discussion is at a general level, consistent with the brevity of this document. Background documents provide further discussion of methods for interpreting clinical evidence.
>
> Second, it proposes specific procedures that the panels should follow in their deliberations. The purpose of these procedures is to ensure that the advice that MCAC panels provide to HCFA is timely and meets the highest standards of comprehensiveness, balance, and scientific quality.

These principles and procedures should make the evaluation process more predictable, more consistent, and more understandable. By making the reasoning behind each panel's conclusions more explicit, these principles should also make the MCAC process more accountable.

HCFA is formulating a proposed rule to outline coverage criteria. The following recommendations are provisional and are meant to assist the Panels in their deliberations until HCFA issues further guidance. We will modify these recommendations as needed to respond to the HCFA final rule about the definition and application of the concept of "reasonable and necessary."

2. EVALUATION OF EVIDENCE

This process is intended to serve the public by identifying medical goods and services that improve the health of Medicare beneficiaries. In advising HCFA about the evidence that a new medical item or service is effective, MCAC panels will need to answer two questions. First, "is the evidence concerning effectiveness in the Medicare population adequate to draw conclusions about magnitude of effectiveness relative to other items or services?" Second, "how does the magnitude of effectiveness of the new medical item or service compare to other available interventions?"

The MCAC panels should explore many sources of evidence in assembling the body of evidence to be used in their deliberations. The sources might include the peer-reviewed scientific literature, the recommendations of expert panels, and unpublished data used to secure FDA approval. The quality of the evidence from these sources will vary, and the panels should weigh the evidence according to its quality.

A. <u>Adequacy of evidence</u>

> **The Panels must determine whether the scientific evidence is adequate to draw conclusions about the effectiveness of the intervention in routine clinical use in the population of Medicare beneficiaries.**

Comment: Assessing the adequacy of the evidence is a *sine qua non* of essentially all modern approaches to the evaluation of medical technologies. Defining what constitutes adequate evidence is a critical step. The committee's definition of adequate evidence includes the *validity* of the evidence and its *general applicability* to the population of interest.

Many forms of evidence can be valid, or not, depending on circumstances specific to the individual study. The most rigorous type of evidence is ordinarily a large, well-designed randomized controlled clinical trial. . The ideal randomized clinical trial has appropriate endpoints, enrolls a representative sample of patients, is conducted in clinical practice in the patient population of interest, and evaluate interventions (diagnostic tests, surgical procedures, medical devices, drugs) as typically used in routine clinical practice.

When several such well-designed trials yield consistent results, there is likely to be a strong consensus that the evidence is sufficient. This level of evidence will likely be unavailable for many of the interventions that the MCAC panels will evaluate. There may be randomized trials conducted in other populations (e.g., middle-aged men rather than men and women 65 years of age and older), randomized trials with important design flaws (e.g., they are not double-blinded), or non-randomized studies with concurrent controls. Deciding whether such studies constitute valid, applicable evidence can be very difficult.

The Executive Committee believes that general guidelines for deciding whether the evidence is adequate will serve our purposes better than a rigid set of standards. In considering the evidence from any study, the MCAC panels should try to answer two main questions:

Bias: Does the study systematically over- or underestimate the effect of the intervention because of possible bias or other errors in assigning patients to intervention and control groups?

There are many potential sources of bias. In observational study designs, the investigators simply observe patient care without intervening to allocate patients to intervention or control groups. In such studies, the investigators cannot be sure that they have measured all of the ways in which treated patients differ from untreated patients. If some of these unmeasured characteristics influence both health outcomes and the likelihood of receiving the intervention, at least part of the measured treatment effect will be a result of the unmeasured patient characteristics rather than the treatment itself. This particular bias is called selection bias. For example, in comparing a new, extensive surgical procedure to a less extensive operation, researchers might measure survival one year after the two procedures. Surgeons might avoid performing an extensive operation on patients with severe comorbid illness. If, in an observational study, the researchers failed to measure comorbid conditions, they might conclude that the patient groups were similar. If patients who got surgery for a disease had a better one year survival rate than those who did not get surgery, the reason could be the good health of those that the surgeons selected for surgery, rather than the surgery itself.

Random allocation of patients to the intervention under study eliminates systematic selection bias. In a properly designed and conducted randomized trial, apart from random differences, the group of patients receiving the intervention and the group receiving the alternative are identical with respect to all characteristics, measured and unmeasured. The investigators can be fairly certain that any observed difference in health outcomes is the result of the intervention. Unbalanced allocation can occur with randomized allocation of subjects, but it is very unlikely when the study groups contain a large number of patients.

In an observational, non-randomized study, it is usually very difficult to determine whether bias could account for the results. However, there may be important exceptions, especially if the intervention dramatically improves the outcome of a disease. For example, if a disease is uniformly fatal within six weeks, and an observational study demonstrates that half of all patients receiving a new treatment survive for at least a year, it is not necessary to conduct a randomized controlled trial to obtain adequate evidence that the treatment is effective. On the other hand, the effect of treatment on the outcomes of most diseases is less predictable than in this extreme case and depends upon difficult-to-measure aspects of each patient's health. In these diseases, bias can strongly influence the results of observational studies. Bias is especially likely if the intervention under study is dangerous or toxic, because physicians might avoid prescribing it for patients who are particularly likely to suffer ill effects. Clinical trials of treatments for cancers that have an unpredictable natural history, for example, have repeatedly demonstrated that the results of observational studies are misleading.

To detect important bias in observational studies, the Panels will need to carefully consider all of the evidence, including the comprehensiveness of the available data, how physicians selected patients to receive the intervention, and the extent of disease in intervention and control group patients. In some cases, the panel may decide that it cannot draw firm conclusions about effectiveness without randomized trials.

Although a body of evidence consisting only of uncontrolled studies – whether based on anecdotal evidence, testimonials, or case series and disease registries without adequate historical controls – is never adequate, in some cases the panel will determine that observational evidence is sufficient to draw conclusions about effectiveness. When these circumstances apply, the panel must describe possible sources of bias andexplain why it decided that bias does not account for the results.

The second question that MCAC panels must strive to answer concerns the external validity of the evidence.

External validity: Do the results apply to the Medicare population?

Historically, many randomized controlled clinical trials excluded older men and women. An increasing number of randomized trials now include elderly men and women. However, simply enrolling older people in proportion to their number in the general population may not be sufficient to determine whether the results of the trial apply to Medicare patients. If the study has too few elderly participants, it might not have the statistical power to detect a clinically important effect in Medicare patients. Clinical trial populations might also differ from the clinically relevant population of Medicare beneficiaries because the trials exclude individuals who have significant comorbid illness or who take many medications. If the study population in the available trials is not the same as the general population of Medicare beneficiaries who would be candidates to receive the intervention, the Panel must state whether the results of the trials apply to typical Medicare patients and explain its reasoning.

Issues of external validity also apply to the intervention. For a drug or device, the intervention is the same when used in different settings. But other interventions may differ from one site to another. For example, the outcomes of a complex surgical procedure can depend heavily on the skills of the surgeons and other staff caring for the patient. If available trials only include sites where surgeons have the best outcomes, the outcomes might be considerably better than what is possible in typical practice settings. The panel must state whether the results are likely to apply to the general practice setting and explain its reasoning.

The second major criterion for evaluating evidence is the size and direction (more effective, as effective, or less effective) of the health effect that it demonstrates.

> **2. <u>Size of Health Effect:</u> Evidence from well-designed studies (meeting criterion #1 above) must establish how the effectiveness of the new intervention compares to the effectiveness of established services and medical items.**

Comment: If the evidence is adequate to draw conclusions (as defined above), the next question is the size and direction of the effect compared with interventions that are widely used. In evaluating the evidence for an intervention, the panels should help HCFA make coverage decisions by placing the size and direction of effectiveness, *as compared to established services or medical items*, into one of these seven categories:

1) Breakthrough technology: the improvement in health outcomes is so large that the intervention becomes the standard of care.

2) Substantially More effective: The new intervention improves health outcomes by a substantial margin as compared with established services or medical items.

3) More effective: the new intervention improves health outcomes by a significant, albeit small, margin as compared with established services or medical items.

4) As effective but with advantages: the intervention has the same effect on health outcomes as established services or medical items but has some advantages (convenience, rapidity of effect, fewer side effects, other advantages) that some patients will prefer.

5) As effective and with no advantages: the intervention has the same effect on health outcomes as established alternatives but with no advantages.

6) Less effective but with advantages: Although the intervention is less effective than established alternatives (but more effective than doing nothing), it has some advantages (such as convenience, tolerability).

7) Less effective and with no advantages: The intervention is less effective than established alternatives (but more effective than doing nothing) and has no significant advantages.

8) Not effective: The intervention has no effect or has deleterious effects on health outcomes when compared with ”doing nothing” (e.g., treatment with placebo or patient management without the use of a diagnostic test).

C. When the Evidence is Insufficient

HCFA may ask MCAC panels for advice when the evidence is ambiguous, scanty, or of poor quality. In this section, the Executive Committee describes some principles to guide the panels when the evidence is not sufficient to draw a strong conclusion.

When a Panel determines that the evidence is insufficient to draw conclusions about the effectiveness of an intervention, it will not attempt to classify the size of the possible effect. Instead, it will explain the reason for its determination and also form a judgment about:

- the possibility of developing better evidence
- the potential benefits of obtaining better information.

Adequate evidence may be unavailable for these reasons:

1. It is not feasible to apply a definitive study design to the intervention or target condition, for any of several reasons.

    - assembling a large enough sample of patients to study is not feasible because the target condition is rare in the study population.
    - double-blinding or even single-blinding is not feasible because the intervention causes distinctive side-effects or has other characteristics that make the patient, or the treating physician, aware that patient is receiving the active intervention rather than placebo

- the intervention alters important health outcomes but only after a delay of years or even decades

Because a panel can expect that the Executive Committee will closely scrutinize a conclusion that studies are not feasible, it should feel an obligation to provide an in-depth explanation of its reasoning. Common obstacles, such as high cost, the difficulties of organizing a large trial, the expense of the intervention, and difficulties in recruitment, are not a sufficient rationale for deciding that a study is not feasible.

2. Definitive studies are possible but have not been performed

- the technology is relatively new
- the cost of performing study is high, and funding has not been available
- studies have been performed but are not definitive

When a panel determines that definitive studies are possible but have not yet been performed, it should also form a judgment about whether the intervention is particularly promising. In this context, “promising” means:

- there are good reasons to expect that the technology will improve health outcomes substantially;
- an improvement in health outcomes appears likely at minimal risk or cost; or
- the intervention would routinely obviate the need for a more risky or costly diagnostic or therapeutic alternative.

HCFA could deal with the problem of inadequately studied but promising technologies in several ways:

- It might encourage or directly support studies that would provide adequate evidence about the effectiveness of promising technologies by directly supporting research.
- It could approve coverage on a provisional basis. For example, it could cover the technology only when it is used in the context of an approved study. Alternatively, it could cover the technology more generally but re-evaluate the coverage decision after adequate time has passed in which to perform definitive studies.
- It could make a coverage decision based upon the best interpretation of the available evidence. Such an approach would give HCFA the flexibility to cover promising treatments for conditions that are too rare to support definitive study.

Other approaches to forming conclusions when the evidence is insufficient:

Although well-designed randomized trials and observational studies are weighted heavily in most evaluations of clinical interventions, other kinds of evidence are relevant and should receive appropriate consideration. Frequently, for example, there is direct evidence from trials about the effects of a treatment on an intermediate endpoint like blood pressure or cholesterol levels, but decision analytic or epidemiological modeling is needed to determine effects on more global health outcomes like the incidence of strokes or heart attacks. Other relevant information includes guidelines from professional societies and other expert bodies, structured and less formal reviews of the literature, and expert testimony. The Panels could consider such information, as long as it can be used to help answer the questions posed to them. Like other bodies that evaluate health care technologies, the Panels should place greater weight on higher quality studies than on studies whose design is flawed or that are not directly relevant to the questions under consideration. They may consider using a framework for grading evidence, for example, like that of the U.S. Preventive Services Task Force.

D. Proposed Guidelines for Evaluating Diagnostic Tests

When they are asked to evaluate diagnostic tests, MCAC panels can apply criteria that are similar to those used for other health interventions that come before the Medicare Coverage Advisory Committee. The panels will need to determine whether the evidence is adequate to conclude that the diagnostic test improves outcomes and, if the evidence is adequate, to classify the magnitude of the health benefit, when a test is used for a specific purpose.

When more than one application of the test is under consideration, the panels will need to evaluate each application. Although this document refers to diagnostic tests, it is important to recognize that tests have four principal uses in clinical settings and that the comments in this document refer to all four uses.

> *Screening*: screening refers to the use of a test to detect asymptomatic, early disease or a predisposition to disease (i.e., a risk factor such as elevated blood pressure or high blood cholesterol). Typically, the pre-test probability of disease (i.e., the prevalence or probability of disease in the population to be screened) is very low in such individuals. The purpose of screening is either to take action to prevent disease by modifying a risk factor, or to detect and treat disease early. In both cases, screening is presumed to be advantageous because early treatment of disease, or modification of a risk factor, improves health outcomes.
>
> *Diagnosis*: a test is used to make a diagnosis when symptoms, abnormalities on physical examination, or other evidence suggests but does not prove that a disease is present. Making a correct diagnosis improves health outcomes by leading to better clinical decisions about further testing and/or treatment.
>
> *Staging:* a test is used to stage a disease when the diagnosis is known but the extent of disease is not known. Staging is particularly important when the stage of disease, as well as the diagnosis itself, influences management. For example, an early stage cancer might be treated surgically, while the same cancer at a more advanced stage might be treated with chemotherapy alone.
>
> *Monitoring*: in a patient known to have a health condition, a test may be useful for monitoring the disease course or the effect of therapy. A monitoring test helps to evaluate the success of treatment and the need for additional testing or treatment.

Although an effective diagnostic test can reduce the morbidity and mortality of disease by guiding clinical decisions, direct proof of effectiveness is usually unavailable. Few studies have directly measured the effects of a diagnostic or screening test on health outcomes (studies of occult blood testing for colon cancer represent one such exception). Typical studies that evaluate the effectiveness of diagnostic, screening, or monitoring tests focus either on technical characteristics (e.g., does a new radiographic test produce higher resolution images) or effects on accuracy (does it distinguish between patients with and without a disease better than another test).

An improvement in the technical performance of a test can lead to improved diagnostic accuracy. For example, a higher resolution imaging study is more likely to distinguish between normal and abnormal anatomic structures, since it is able to delineate both types of structures more clearly. Improved technical characteristics do not always lead to greater test accuracy and clinical utility. Often, the technical performance of the test is not the factor that limits the ability of a test to distinguish between diseased and non-diseased, or between a person at high risk for disease and a person at average risk. . Sometimes, the indicator that we are trying to measure (e.g., the risk factor) is only imperfectly correlated with the health condition, and improved measurement of the indicator will not lead to greater accuracy. Occasionally, technical performance can improve one aspect of a test's utility while worsening another; for example, MRI scans have higher resolution than most CT scans. Thus MRI scans were initially believed to be superior to CT scans for most indications. However, because CT scans are better able to distinguish certain tissue types, they proved to be better at detecting some abnormalities than the higher-resolution MRI scans. Thus improvements in aspects of technical performance are not sufficient to establish improved diagnostic accuracy.

When good quality studies directly measure how the use of a diagnostic test affects health outcomes, the panel can easily determine that the evidence is adequate and draw conclusions about the magnitude of the health benefits. But when the best studies only measure the accuracy of the test, the panels will have to determine whether the evidence is adequate to conclude that the test improves the accuracy of diagnosis or staging of disease *and* that the improvement in accuracy leads to better health outcomes.

When a panel evaluates a diagnostic test, we suggest that it answer the following question:

> ***Is the evidence adequate to conclude whether the use of the diagnostic test leads to a clinically significant improvement, worsening, or no change in health outcomes, when compared to an alternative clinical strategy?***

The alternative strategy could be, for example, the use of another test, use in combination with another test, or the use of no test at all (e.g., the alternative is treatment or observation without testing). Without evidence to determine whether the test in question leads to a change in health outcomes, it is not possible to make an informed judgment about its appropriate clinical role.

If *direct* evidence linking the use of the test to health outcomes is not available, the panels should answer the following questions, which collectively determine whether there is convincing *indirect* evidence that the test will lead to better health outcomes:

Question 1: *Is the evidence adequate to determine whether the test provides more accurate diagnostic information?*

Question 1 applies when the alternative under consideration is another diagnostic strategy. The definition of "more accurate" is crucial. The standard measures of accuracy are **sensitivity** (probability of a positive test result in a patient with a disease or risk factor or other health condition) and **specificity** (the probability of a negative test result in a patient who does not have the disease). Ideally a new test would increase *both* sensitivity and specificity, but often it does not. A test that has a higher sensitivity is not unambiguously more accurate than an alternative test unless its specificity is at least as great. For most diagnostic tests, a change in the definition of an abnormal result will change the sensitivity, but improved sensitivity is obtained at the cost of worsened specificity, and vice versa. For example, if the diagnosis of diabetes is made on the basis of a fasting blood sugar, the use of a lower blood sugar level to define diabetes results in greater sensitivity but lowered specificity when compared to a diagnostic threshold at a higher blood glucose level. By choosing a different threshold, it is possible to change sensitivity without changing the test. Thus, if only sensitivity (or specificity) were considered, the same test might appear more accurate solely because the definition of an abnormal test result was changed.

The foregoing discussion leads to the following definition of "more accurate:" A more accurate test is not only more sensitive (or specific); it *has a higher sensitivity for a given level of specificity* when compared to another test. At a minimum, then, to conclude that one test is more accurate than another, its sensitivity (or specificity) is must be higher while its specificity (or sensitivity) is the same or better than the alternative test or diagnostic strategy.[1]

In deciding whether one test is more accurate than a second, established test, the panels will need to evaluate the quality of the studies of test performance. In assessing the quality of studies, panels might first consider the characteristics of an "ideal" study of test accuracy and compare the existing studies to the ideal. "Ideal" and "typical" studies of a screening, diagnostic, or monitoring test differ in these ways:

Better study	Typical study	Effect of Typical Study
The study subjects are consecutive patients seen in a typical clinical setting with a chief complaint.	Subjects selected because they had the diagnostic gold standard.	Overestimates sensitivity and underestimates specificity
All patients who get the index test also get the reference test	Patients with negative results on the index test often don't get the diagnostic gold standard	Overestimates sensitivity and underestimates specificity
The person who interprets the index test is blinded to	The person who interprets the index knows the clinical history and the results of	Overestimates sensitivity

all other information	the diagnostic gold standard.	and specificity.
The person who interprets the reference test is blinded to all other information	The person who interprets the diagnostic gold standard knows the clinical history and the results of the index test.	Overestimates sensitivity and specificity.
The reference test is a valid measure of the disease state	The diagnostic gold standard imperfectly measures the disease state.	The measured test performance could either be worse or better than the true performance.

*The **reference test** is a test that is considered the "gold standard," i.e., a test that is used to define the disease. Tests commonly used as reference tests are coronary angiography, for coronary artery disease, and histopathology, for cancer. Reference test can be interpreted more broadly to mean any method that is considered the definite basis for determining whether a disease or risk factor is truly present.

The panels will need to decide whether the estimated accuracy of a test in a study that falls short of the ideal is likely to be distorted by a substantial degree of bias, or whether the limitations of the study are sufficiently minor that it is possible to draw conclusions about the accuracy of the test.

Often an important question is whether the test under consideration complements another test by detecting patients that the first test does not detect. Although the relevant comparison is often between the test under consideration and an alternative test, the comparison will sometimes be between doing a second, additional diagnostic test and not doing an additional test. For example, the question may be whether to do another imaging procedure (e.g. PET scan) when an imaging procedure (e.g., CT scan) has already been done. In this context, the sensitivity and specificity of a new test can be the same as – or even worse than – the sensitivity and specificity of an established comparison test, yet still provide valuable information. It can add value if it provides *complementary* information. In this circumstance, a combination of the two tests leads to more accurate distinction between patients with and without the disease (or risk factor) than either test individually. The information is likely to be complementary if the new test or tests detect other features of the disease (for example, one test measures a physiological phenomenon while the other is an imaging test that detects structural abnormalities). To compare strategies using the two tests and those using only the standard test, one can study patients who receive both tests as well as the reference test (or any direct measure of whether disease is actually present). The appendix describes how such a study can be used to determine whether the combined testing strategy improves the accuracy of diagnosis.

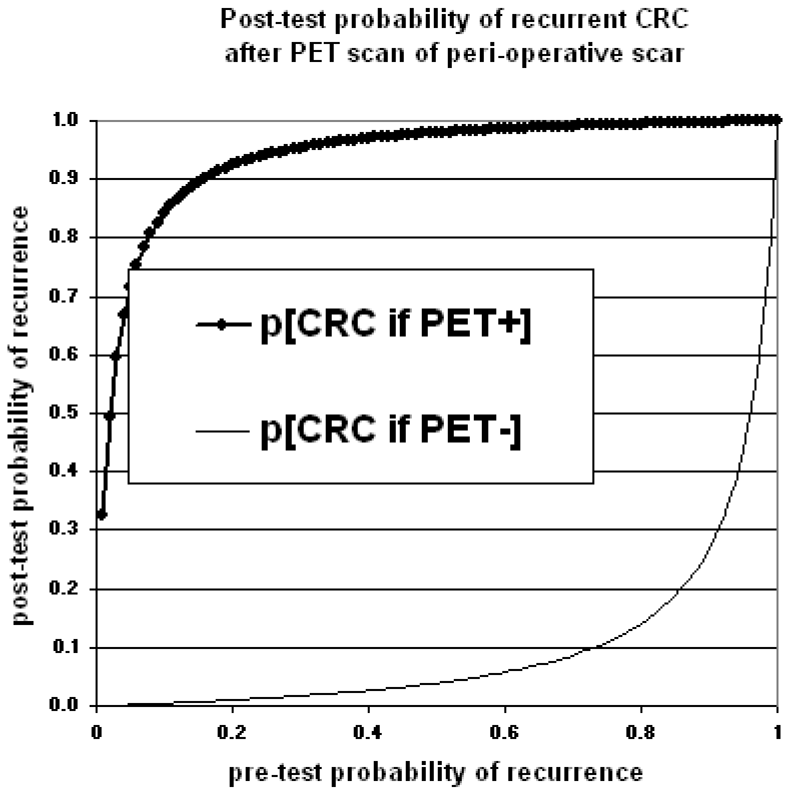
Question 2: ***If the test changes accuracy, is the evidence adequate to determine how the changed accuracy affects health outcomes?***

To determine whether a difference in test accuracy would lead to important changes in health outcomes, the panels may find the following steps helpful.

Step 1: Calculate the post-test probability of disease

The purpose of testing is to reduce uncertainty about the presence of a disease or risk factor, or about the extent of a previously diagnosed disease. The pre-test probability of disease is the probability of disease, risk factor, or extent of disease before the test has been performed, based upon history, physical examination, and preliminary diagnostic tests. The pre-test probability is often used interchangeably with the term “disease prevalence,” but the two terms are only equivalent when prevalence and pre-test probability are based on the same population (i.e, adjusted for history and other information).

The post-test probability is the probability of disease after learning the test results. A test result should only change patient management if it changes the probability of disease. Bayes’ theorem is the formal approach used to calculate the post-test probability. Application of Bayes’ theorem in this context requires knowing the sensitivity and specificity of the test and the pre-test probability of disease. Generally, tests have the greatest effect on probability (i.e., in comparison to the pre-test probability) when the pre-test probability is intermediate (i.e., not near a probability of either 0 or 1). Conversely, tests alter probability the least when the pre-test probability is close to zero or close to 1.0. Often, the patient’s symptoms, abnormalities on physical examination, and other evidence strongly suggest that the patient has the disease in question (i.e., the pre-test probability of disease is high). Unless a test is extremely sensitive, the a patient with a very high pre-test probability is likely to have the disease even if the test result is negative, and should be managed accordingly. Similarly, if the pre-test risk of disease is very low, the probability of disease in a patient with a positive test result remains very low, unless the test is extremely specific (i.e., rarely produces false-positive results). The accompanying graph of post-test probability for two tests illustrates this point. Panels may find these graphs helpful in interpreting the possible impact of a difference in test performance.



The same principles apply to the use of testing to establish the stage of a disease or to monitor the effect of treatment. In these situations, the uncertainty is not about the diagnosis. Rather, the test reduces uncertainty about the current status of the disease. Learning more about stage or response to treatment is important insofar as it will influence a management decision – for example, disease progression while on one treatment will often lead to a change in therapies or cessation of a potentially toxic therapy. A false-negative staging test result (i.e., one that implies the disease is more limited than it really is) may lead to treatment that is both ineffective and harmful. In some situations, a false-positive staging test result can have even more harmful consequences; the physician could withhold potentially curative treatment if he or she interprets the staging test as indicating that cure is not possible, dooming a patient to die of a disease that could have been treated effectively.

Step 2: Evaluate the potential impact on management when alternative tests lead to different post-test probabilities of disease*:*

In the absence of direct evidence of the effects of a test on health outcomes, it will sometimes be possible to conclude with great confidence that improved accuracy will lead to better outcomes. This conclusion is particularly likely when the treatment or management strategy is effective for patients with the disease, but poses risks or discomfort that would not be acceptable when administered to patients who do not have the disease. Then, improved accuracy leads to effective treatment for more people who truly have the disease, while helping to avoid unnecessary treatment in people who would not benefit from it. Thus, although the evidence that diagnostic tests for cancer and for heart disease alter health outcomes is largely indirect, it is often compelling. For these categories of disease, there is often strong evidence that treatments with significant adverse consequences are effective when used appropriately. Panels will need to judge whether the test leads to better patient management by increasing the rate at which patients with disease receive appropriate treatment while reducing the rate at which patients who do not have the disease receive unnecessary treatment.

If management changes, the improvement in health outcomes should be large enough to convince the panel that it is clinically significant. A small increase in accuracy can lead to substantial improvements in health outcomes if treatment is highly effective. Improved accuracy is of little consequence, however, if treatment is either ineffective, so there is little benefit to patients with the disease, or very safe, so there is little harm to patients without the disease. When a treatment has little effect on anyone, improved accuracy is unlikely to lead to improved health outcomes or even to influence clinical decisions.

Under exceptional circumstances, prognostic information, even if it did not affect a treatment decision, could improve health outcomes by improving a patient's sense of well-being. The panel should be alert for circumstances in which patients would be likely to value prognostic informationso much that the information would significantly alter their well-being.

Summary

The recommended approach for evaluating diagnostic tests is as follows:

- Review, when available, high quality studies that provide *direct* evidence that test results improve health outcomes.
- If there is no high quality *direct* evidence, evaluate the *indirect* evidence as follows:

  Decide whether studies of test accuracy are sufficiently free of bias to permit conclusions about the accuracy of the test under consideration, in comparison either to another test or another screening, diagnostic, or staging strategy

  Evaluate the potential impact of improved accuracy (or complementary information) on health outcomes. Evaluating the effect of test accuracy on post-test probability is one part of this step. The other part is deciding whether the change in patient management that results from the test will improve health outcomes. Improved outcomes are most likely to occur when the management strategy is effective in patients with the disease and does not benefit or even harms those without the disease. Thus, a test can improve health outcomes when the treatment poses such a significant risk of harm that it is very important to avoid unnecessary treatment.

[1] The more technical expression of this condition is that a more accurate test is one whose receiver operating characteristic (ROC) curve is above and to the left of the ROC curve for the alternative test.